J. R. Statist. Soc. B (1995)
57, *No.* 2, *pp.* 425–437

A Gradient Algorithm Locally Equivalent to the EM Algorithm

By KENNETH LANGE†

University of Michigan, Ann Arbor, USA

[Received March 1992. Final Revision May 1994]

SUMMARY

In many problems of maximum likelihood estimation, it is impossible to carry out either the E-step or the M-step of the EM algorithm. The present paper introduces a gradient algorithm that is closely related to the EM algorithm. This EM gradient algorithm approximately solves the M-step of the EM algorithm by one iteration of Newton's method. Since Newton's method converges quickly, the local properties of the EM gradient algorithm are almost identical with those of the EM algorithm. Any strict local maximum point of the observed likelihood locally attracts the EM and EM gradient algorithm at the same rate of convergence, and near the maximum point the EM gradient algorithm always produces an increase in the likelihood. With proper modification the EM gradient algorithm also exhibits global convergence properties that are similar to those of the EM algorithm. Our proof of global convergence applies and improves existing theory for the EM algorithm. These theoretical points are reinforced by a discussion of three realistic examples illustrating how the EM gradient algorithm can succeed where the EM algorithm is intractable.

Keywords: CONVERGENCE; DIRICHLET DISTRIBUTION; MAXIMUM LIKELIHOOD; ROBUST REGRESSION; SURVIVAL ANALYSIS

1. INTRODUCTION

The EM algorithm is one of the most versatile algorithms in modern statistics (Dempster *et al.*, 1977; Little and Rubin, 1987). Because of its simplicity and numerical stability, it is often the method of choice for computing maximum likelihood or maximum *a posteriori* estimates. In truth, the EM algorithm is not so much an algorithm as a prescription for an algorithm. In many interesting examples, either the E-step or the M-step proves intractable. When faced with such a dilemma, most statisticians immediately turn to other algorithms such as scoring. However, statisticians are beginning to devise tactics to overcome the lack of explicit solutions of either the E-step or the M-step. Part of the motivation for the Gibbs sampler and data augmentation certainly originates with frustration in solving the E-step (Wei and Tanner, 1990).

Intractability of the M-step has also begun to yield to new numerical techniques. For instance, in the ECM algorithm of Meng and Rubin (1993), the M-step is done in cyclic fashion with different subsets of the parameters successively incremented. In some real examples, it is possible to solve the M-step for subsets of parameters but not for all parameters simultaneously. The one-step-late (OSL) algorithm of Green (1990) is designed for maximum *a posteriori* estimation. Green suggests this

†*Address for correspondence*: Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA.
E-mail: klange@umich.edu

 0035–9246/95/57425

algorithm when the introduction of a prior renders the M-step intractable. In specific applications the OSL algorithm is derived by temporarily fixing the gradient of the log-prior and then solving the M-step with this qualification. Finally, the gradient algorithm of Titterington (Titterington, 1984; Titterington *et al.*, 1985) mimics the EM algorithm for linear exponential families.

In the present paper, we explore a variation of the EM algorithm that is closest in form to Titterington's algorithm. This new gradient algorithm has convergence properties that are almost identical with those of the standard EM algorithm. In fact, we shall show that it has the same local rate of convergence and, with proper precautions, the same desirable property of always leading uphill. To explain this new algorithm, let us recall the conventions underlying the EM algorithm (Dempster *et al.*, 1977; Little and Rubin, 1987). An important distinction is drawn between the observed incomplete data Y and the unobserved complete data X. The complete data X are assumed to have probability density $f(X|\theta)$, which is a function of a parameter vector θ as well as of X. In the E-step of the EM algorithm, the conditional expectation

$$Q(\theta|\theta^n) = E[\ln\{f(X|\theta)\}|\, Y,\ \theta^n]$$

is computed. Here θ^n is the current estimated value of θ. In the M-step, the θ maximizing $Q(\theta|\theta^n)$ is found. This yields the new parameter estimate θ^{n+1}, and this two-step process is repeated until convergence. The essence of the EM algorithm is that increasing $Q(\theta|\theta^n)$ forces an increase in the log-likelihood $L(\theta)$ of the observed data.

If it is impossible to carry out the M-step exactly, we can contemplate solving it iteratively. This is an unattractive alternative because it involves iterating within iterations. The fastest common algorithm for iteratively solving the M-step would be Newton's method, which has quadratic convergence compared with the linear convergence experienced in the EM algorithm. These considerations suggest that, perhaps, a single iteration of Newton's method at each M-step would be adequate to ensure convergence of an approximate EM algorithm. This heuristic argument forms the basis of our new gradient algorithm. Update the current parameter column vector θ^n by

$$\begin{aligned}\theta^{n+1} &= \theta^n - \mathrm{d}^{20}Q(\theta^n|\theta^n)^{-1}\,\mathrm{d}^{10}Q(\theta^n|\theta^n)\\ &= \theta^n - \mathrm{d}^{20}Q(\theta^n|\theta^n)^{-1}\,\mathrm{d}L(\theta^n). \qquad (1)\end{aligned}$$

In equation (1) the operators d^{10} and d^{20} take first and second partial derivatives respectively with respect to the first variable of Q. The column vector $\mathrm{d}L(\theta)$ is the score of the log-likelihood $L(\theta)$. Because $L(\theta) - Q(\theta|\theta^n)$ has its minimum at $\theta = \theta^n$, the equality $\mathrm{d}^{10}Q(\theta^n|\theta^n) = \mathrm{d}L(\theta^n)$ holds whenever θ^n is an interior point of the parameter domain (Dempster *et al.*, 1977). We shall refer to algorithm (1) as the EM gradient algorithm.

In his gradient algorithm, Titterington (1984) substituted in equation (1) the Fisher information matrix of the complete data X for the matrix $-\mathrm{d}^{20}Q(\theta^n|\theta^n)$. When the complete data belong to a linear exponential family, the two matrices coincide. One advantage of Titterington's algorithm is that it is necessarily an ascent algorithm. This means that a fractional step in the current direction will certainly lead to an increase in $L(\theta)$. We can retain this advantage in the EM gradient

algorithm by demanding that $d^{20}Q(\theta^n|\theta^n)$ be negative definite. In all the examples discussed later, the Hessian matrix $d^{20}Q(\theta|\theta^n)$ is indeed always negative definite. This fact in turn implies strict concavity of $Q(\theta|\theta^n)$ and uniqueness of the maximum point θ^{n+1}. In some cases negative definiteness of $d^{20}Q(\theta^n|\theta^n)$ can only be achieved by reparameterization; this does not change the EM algorithm but does affect the EM gradient algorithm.

In the next section, we present some specific examples where the M-step of the EM algorithm is intractable, but the EM gradient algorithm is straightforward. The third and fourth sections of the current paper are devoted to a theoretical development of the EM gradient algorithm. This gives us a chance to verify some of the claims made for the algorithm and, in particular, to investigate its convergence properties. Some of our results improve existing theory for the EM algorithm as well (Dempster *et al.*, 1977; Boyles, 1983; Wu, 1983). In the concluding discussion, we note some generalizations and limitations of the EM gradient algorithm.

2. EXAMPLES

2.1. *Dirichlet Distribution*

First consider an example where the complete data belong to a linear exponential family. Suppose that $X_1, \ldots, X_k$ are independent random variables with X_i having gamma density

$$\Gamma(\theta_i)^{-1}x_i^{\theta_i-1}\exp(-x_i)$$

for $x_i > 0$. With the superscript $*$ denoting vector transpose, the Dirichlet random vector $Y = (Y_1, \ldots, Y_k)^*$ is defined by setting its ith component equal to the proportion

$$Y_i = X_i \Big/ \sum_{j=1}^{k} X_j.$$

It can be shown that Y has density

$$g(y|\theta) = \frac{\Gamma\left(\sum_{i=1}^{k}\theta_i\right)}{\prod_{i=1}^{k}\Gamma(\theta_i)} \prod_{i=1}^{k} y_i^{\theta_i-1}$$

on the simplex $\{y = (y_1, \ldots, y_k)^*: y_1 > 0, \ldots, y_k > 0, \Sigma_{i=1}^k y_i = 1\}$ endowed with the uniform measure. The random vector Y constitutes the observed data, and the underlying random vector $X = (X_1, \ldots, X_k)^*$ constitutes the complete data.

For an independent and identically distributed sample $Y^1, \ldots, Y^m$ from the Dirichlet distribution, we can attempt to estimate the parameter vector $\theta = (\theta_1, \ldots, \theta_k)^*$ by the EM algorithm. Let $X^1, \ldots, X^m$ be the corresponding complete data. It is immediately evident that up to an irrelevant constant

$$Q(\theta|\theta^n) = -m\sum_{i=1}^{k}\ln\Gamma(\theta_i) + \sum_{i=1}^{k}(\theta_i - 1)\sum_{j=1}^{m}E(\ln X_i^j | Y^j, \theta^n). \tag{2}$$

Owing to the presence of the terms $\ln\Gamma(\theta_i)$ in equation (2), we cannot solve

the M-step analytically. However, the EM gradient algorithm is trivial to implement since the Hessian matrix $d^2Q(\theta|\theta^n)$ is diagonal with ith diagonal entry $-m\, d^2\{\ln\Gamma(\theta_i)\}/d\theta_i^2$. In this example as in other linear exponential examples, it is unnecessary to evaluate the conditional expectations of the E-step.

Scoring is an attractive alternative to the EM gradient algorithm in this particular example (Narayanan, 1991). The data of Mosimann (1962) on the relative frequencies of $k = 3$ serum proteins in $m = 23$ young, white Pekin ducklings furnish an interesting test case for comparing the EM gradient algorithm with scoring. Starting from $\theta^1 = (1., 1., 1.)^*$, both algorithms converge to the maximum point $(3.22, 20.38, 21.69)^*$ with the log-likelihood *en route* showing a steady increase to its maximum value of 73.12500. However, scoring takes only nine iterations for the log-likelihood to achieve its final value, whereas the EM gradient algorithm takes 333 iterations. This example just confirms the relatively slow convergence of the EM algorithm. Speed of convergence must always be balanced against numerical stability and ease of programming. Here the EM gradient algorithm avoids matrix inversion. Both algorithms need to face the unpleasant task of evaluating log-gamma, digamma and trigamma functions (Pike and Hill, 1966; Schneider, 1978).

2.2. *Adaptive Robust Linear Regression via t-distribution*

Lange *et al.* (1989) investigated an EM algorithm for adaptive robust regression based on t-distributed errors (Dempster *et al.*, 1980). A t-distributed random variable Y with mean μ, scale σ and degrees of freedom ν can be represented by the shifted ratio $Y = \mu + V/\sqrt{U}$, where $U = \chi_\nu^2/\nu$ is a χ^2 random variable scaled to have mean 1 and V is an independent normal random variable with mean 0 and standard deviation σ. The pair (Y, U) constitutes the complete data corresponding to the observed data Y. We shall let the mean $\mu(\zeta)$ depend on a parameter vector ζ and denote the concatenated parameter vector $(\zeta^*, \sigma, \nu)^*$ by θ.

If we observe a sequence of independent observations $Y^1, \ldots, Y^m$ having different mean functions $\mu^1(\zeta), \ldots, \mu^m(\zeta)$ but the same scale σ and degrees of freedom ν, then the E-step of the EM algorithm gives up to a constant

$$Q(\theta|\theta^n) = -m\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^m w^i\{Y^i - \mu^i(\zeta)\}^2 - m\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{m\nu}{2}\ln\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2} - 1\right)\sum_{i=1}^m q^i - \frac{\nu}{2}\sum_{i=1}^m w^i,$$

where $w^i = E(U^i|Y^i, \theta^n)$ and $q^i = E(\ln U^i|Y^i, \theta^n)$ are given by $(\nu^n + 1)/\{\nu^n + d^i(\theta^n)\}$ and $DG\{(\nu^n + 1)/2\} - \ln[\{\nu^n + d^i(\theta^n)\}/2]$ respectively. Here $d^i(\theta^n) = [\{Y^i - \mu^i(\zeta^n)\}/\sigma^n]^2$ and $DG(s)$ denotes the digamma function $d\{\ln\Gamma(s)\}/ds$ (Lange *et al.*, 1989).

The M-step of the EM algorithm simplifies because maximization over ζ and σ separates from maximization over ν. When each $\mu^i(\zeta) = \zeta^* z^i$ is a linear function of a covariate vector z^i, then maximization over ζ and σ reduces to weighted linear regression. Maximization over ν cannot be done in closed form and requires an iterative solution. The EM gradient algorithm avoids the inner iteration on ν, but it may encounter difficulties since $d^2Q(\theta^n|\theta^n)$ is not necessarily negative definite. This problem can be circumvented by the reparameterization $\alpha = 1/\sigma$ and $\beta = \zeta/\sigma$

(Pratt, 1981). Retaining the symbol θ for the concatenated parameter vector $(\beta^*, \alpha, \nu)^*$, it is straightforward to check that $d^2Q(\theta^n|\theta^n)$ is block diagonal with upper left block

$$-m\begin{pmatrix}0 & 0\\ 0 & \alpha^{-2}\end{pmatrix} - \sum_{i=1}^{m} w^i \begin{pmatrix}-z^i\\ Y^i\end{pmatrix}(-(z^i)^* \quad Y^i)$$

corresponding to β and α and lower right block $m/2\nu - (m/4)\,\mathrm{TG}(\nu/2)$ corresponding to ν, where $\mathrm{TG}(s)$ denotes the trigamma function (Hille, 1959). The upper left block is negative definite provided that the matrix $(z^1 \ldots z^m)$ has rank equal to the number of components of β. The lower-right block is negative because the trigamma function satisfies

$$\begin{aligned}\frac{1}{2\nu} - \frac{1}{4}\mathrm{TG}\left(\frac{\nu}{2}\right) &= \frac{1}{2\nu} - \frac{1}{4}\sum_{k=0}^{\infty}\frac{1}{(\nu/2+k)^2}\\ &< \frac{1}{2\nu} - \frac{1}{4}\int_0^{\infty}\frac{1}{(\nu/2+s)^2}\,ds\\ &= 0.\end{aligned}$$

Lange and Sinsheimer (1993) gave a numerical example of robust linear regression with two mean parameters. After the reparameterization $\theta = (\beta^*, \alpha, \nu)^*$, the EM algorithm takes 51 iterations on this problem, and the EM gradient algorithm takes 53 iterations. However, the EM algorithm requires two or three inner iterations per M-step to compute the update of ν. Thus, the EM gradient algorithm has an edge in computational speed. Titterington's algorithm can also be used for this problem. The Fisher information matrix for the complete data is block diagonal, having upper left block

$$m\begin{pmatrix}0 & 0\\ 0 & 2\alpha^{-2}\end{pmatrix} + \sum_{i=1}^{m}\begin{pmatrix}-z^i\\ \beta^* z^i\end{pmatrix}(-(z^i)^* \quad \beta^* z^i)$$

and lower right block $-m/2\nu + (m/4)\,\mathrm{TG}(\nu/2)$. Titterington's algorithm takes a painful 5420 iterations to reach the maximum likelihood of -119.5427. *En route* many of the computed increments diminish the log-likelihood, and the standard remedy of step halving must be employed as a countermeasure.

2.3. *Markov Chain Model for Survival Analysis*

Consider a continuous time Markov chain $\{Z(t), t \geqslant 0\}$ with k states and with infinitesimal transition rates λ_{ij} between pairs of states i and j. The finite time transition probabilities $P_{ij}(t) = \Pr(Z_t = j | Z_0 = i)$ can be collectively expressed by the matrix exponential $P(t) = \exp(t\Lambda)$, where the matrix Λ has off-diagonal entries $\Lambda_{ij} = \lambda_{ij}$ and diagonal entries $\Lambda_{ii} = -\lambda_i = -\Sigma_{j\neq i}\lambda_{ij}$ (Chiang, 1980).

In modelling the effects of covariates on the transition rates, it is convenient to adopt the functional form $\lambda_{ij} = \exp(\theta^* w^{ij})$, where w^{ij} is a vector of covariates appropriate to transitions between states i and j. In survival analysis models of cancer progression, the disease state of a person in a clinical study is observed at some sequence of times $t_1 < \ldots < t_m$. The covariates may change from one time

interval (t_r, t_{r+1}) to the next (t_{r+1}, t_{r+2}), but within a time interval we assume that the covariates, and hence the transition rates as well, are constant.

The natural complete data corresponding to the observed sequence of states $Z_{t_1} = i_1, \ldots, Z_{t_m} = i_m$ on a patient is the whole process $\{Z(t): t_1 \leqslant t \leqslant t_m\}$. The Markov assumption implies the patient's likelihood factors as

$$\Pr(Z_{t_1} = i_1) \prod_{j=2}^{m} \Pr(Z_{t_j} = i_j | Z_{t_{j-1}} = i_{j-1}). \tag{3}$$

In general, it is preferable to condition on the initial state i_1 at time t_1 and to employ a conditional likelihood omitting the leftmost factor of expression (3). This amended representation makes it clear that we can view the evolution of the patient's history as a sequence of multinomial trials. Without loss of generality, we can therefore focus on a single patient who starts for convenience in state i at time 0 and ends in state j at time t.

To deal with the interval censoring of the complete data, suppose that the Markov chain is in state r at the moment of some transition. The neighbouring state s is chosen with probability λ_{rs}/λ_r. This probability must be multiplied by the density $\lambda_r \exp(-\lambda_r u)$, assuming that the chain has spent an amount of time u in state r since last arriving there. Thus, this transition and its previous waiting period together contribute a factor of $\lambda_{rs} \exp(-\lambda_r u)$ to the complete data likelihood. These contributions should be supplemented by the factor $\exp(-\lambda_j v)$ for the duration v of the stay in the final state j after the last transition during $[0, t]$. The complete data log-likelihood can therefore be expressed as

$$\sum_{(r,s)} N_{rs} \ln \lambda_{rs} - \sum_r T_r \lambda_r,$$

where N_{rs} is the random number of transitions from r to s during $[0, t]$ and T_r is the random length of time in state r during $[0, t]$.

To implement the E-step of the EM gradient algorithm, we must compute the conditional expectations of each N_{rs} and T_r given the initial and final states. Fortunately, these conditional expectations can be expressed in terms of the finite time transition probabilities as

$$E(N_{rs} | Z_0 = i, Z_t = j) = \frac{\int_0^t P_{ir}(u) \lambda_{rs} P_{sj}(t-u)\, du}{P_{ij}(t)}, \tag{4}$$

$$E(T_r | Z_0 = i, Z_t = j) = \frac{\int_0^t P_{ir}(u) P_{rj}(t-u)\, du}{P_{ij}(t)}. \tag{5}$$

To verify expression (4) consider a small time interval $(u, u + du)$. On this interval there is a transition from r to s with probability $\lambda_{rs}\, du$, provided that the chain is poised in state r at time u. Once the transition takes place, the chain must move on to state j from state s in the remaining $t - u$ units of time. The denominator in expression (4) arises from conditioning on the initial and final states. Verification of expression (5) follows by similar reasoning.

Expressions (4) and (5) can be easily evaluated since each entry of $P(t)$ is typically

a linear combination of exponential functions. If we now suppose that $P_{ir}(t) = \sum_{m=1}^{k} c_m \exp(\rho_m t)$ and $P_{sj}(t) = \sum_{n=1}^{k} d_n \exp(\rho_n t)$, then

$$\int_0^t P_{ir}(u)\lambda_{rs} P_{sj}(t-u)\,du = \lambda_{rs} \sum_{m=1}^{k} \sum_{n \neq m} c_m d_n \frac{\exp(\rho_m t) - \exp(\rho_n t)}{\rho_m - \rho_n} + \lambda_{rs} \sum_{m=1}^{k} c_m d_m \exp(\rho_m t) t. \tag{6}$$

Although for this model the M-step of the EM algorithm is thwarted by an intractable system of transcendental equations, the EM gradient algorithm behaves well. Negative definiteness of $d^{20}Q(\theta^n \mid \theta^n)$ follows immediately from its representation

$$-\sum_r E(T_r \mid Z_0 = i,\ Z_t = j,\ \theta^n) \sum_{s \neq r} w^{rs}(w^{rs})^* \exp\{(\theta^n)^* w^{rs}\}$$

for a single pair of adjacent times on a single patient under the parameterization $\lambda_{rs} = \exp(\theta^* w^{rs})$.

Wanek *et al.* (1993) carried out a detailed analysis of melanoma data according to the above Markov chain model. They obtained maximum likelihood estimates by a version of the ECM algorithm of Meng and Rubin (1993) that depends on numerically maximizing the Q-function one parameter at a time. A preliminary comparison of this algorithm and the EM gradient algorithm suggests that the EM gradient algorithm is at least an order of magnitude faster.

3. LOCAL CONVERGENCE OF ALGORITHM

Not surprisingly, the theoretical properties of the EM gradient algorithm closely parallel those of the EM algorithm. To keep our discussion reasonably short, we shall make several simplifying assumptions. First, it helps to restrict attention to examples where the parameter domain U is an open convex set of some Euclidean space $\mathscr{R}^k$. Let θ, and occasionally ϕ and ψ, denote typical elements of U. We shall require that the log-likelihood $L(\theta)$ be continuous and upper compact in the sense that the set $\{\theta \in U\colon L(\theta) \geq c\}$ is compact for every constant c. Upper compactness implies that $L(\theta)$ tends to $-\infty$ as θ approaches the boundary of U and that $L(\theta)$ has at least one maximum point. It is also convenient to suppose that $L(\theta)$ and $Q(\theta \mid \phi)$ and their first and second differentials with respect to θ are jointly continuous in θ and ϕ. Finally, as in the examples, we demand that $d^{20}Q(\theta \mid \phi)$ be negative definite. Among other things, this implies that the matrix inverse $d^{20}Q(\theta \mid \theta)^{-1}$ always exists and that the iterates of the algorithm are well defined, except possibly for the question of whether they fall in U.

With these provisos, we investigate the local convergence properties of the EM gradient algorithm (1). For brevity call the EM gradient algorithm map $M(\theta)$. Now suppose that θ^∞ is a stationary point of $L(\theta)$. Proceeding formally, the differential $dM(\theta^\infty)$ should be

$$\begin{aligned} dM(\theta^\infty) &= I - d^{20}Q(\theta^\infty \mid \theta^\infty)^{-1} d^2L(\theta^\infty) + d\{d^{20}Q(\theta^\infty \mid \theta^\infty)^{-1}\} dL(\theta^\infty) \\ &= I - d^{20}Q(\theta^\infty \mid \theta^\infty)^{-1} d^2L(\theta^\infty) \\ &= d^{20}Q(\theta^\infty \mid \theta^\infty)^{-1}\{d^{20}Q(\theta^\infty \mid \theta^\infty) - d^2L(\theta^\infty)\}. \end{aligned} \tag{7}$$

The fact that $dL(\theta^\infty) = 0$ saves us the trouble of evaluating the differential of $d^{20}Q(\theta|\theta)^{-1}$ at θ^∞. A careful proof of formula (7) assuming no third derivatives of $Q(\theta|\theta)$ is given in Ortega (1990). In any case, the differential (7) coincides with the differential of the EM algorithm map (Dempster *et al.*, 1977).

Proposition 1. Under the above assumptions, if θ^∞ is a local maximum of $L(\theta)$ such that $d^2L(\theta^\infty)$ is negative definite, then the EM gradient algorithm is locally attracted to θ^∞. The linear rate of convergence to θ^∞ is determined by the dominant eigenvalue of $dM(\theta^\infty)$.

Proof. Because the matrix $-d^{20}Q(\theta^\infty|\theta^\infty)$ is positive definite and the matrix difference $d^2L(\theta^\infty) - d^{20}Q(\theta^\infty|\theta^\infty)$ is positive semidefinite, the theory of relative eigenvalues for symmetric matrices (Hestenes, 1981) implies that all eigenvalues of $dM(\theta^\infty)$ lie on the half open interval $[0, 1)$. The proposition is therefore an immediate consequence of Ostrowski's theorem (Ortega, 1990). □

The slow convergence of the EM and EM gradient algorithms can be ameliorated by inflating the current step by a fixed factor (Redner and Walker, 1984). Consider the modified EM gradient map $M_t(\theta) = \theta + t\{M(\theta) - \theta\}$ for $t > 0$. At θ^∞ this map has differential $dM_t(\theta^\infty) = (1 - t)I + t\,dM(\theta^\infty)$. To every eigenvalue ω of $dM(\theta^\infty)$ there corresponds an eigenvalue $\omega_t = 1 - t + t\omega$ of $dM_t(\theta^\infty)$ and vice versa. Because every $\omega \in [0, 1)$, it is easy to deduce that every $\omega_t \in (-1, 1)$ when $0 < t < 2$. In this range of t, the spectral radius of $dM_t(\theta^\infty)$ is less than 1, and Ostrowski's theorem again implies the local attraction of $M_t(\theta)$ to θ^∞.

It is noteworthy that there is an optimal choice of t. Following the traditional development of successive relaxation in linear algebra (Hämmerlin and Hoffmann, 1991), suppose that the largest and smallest eigenvalues of $dM(\theta^\infty)$ are $\omega_{\max}$ and $\omega_{\min}$. Then the spectral radius of $dM_t(\theta^\infty)$ is given by $\max\{|1 - t + t\omega_{\min}|, |1 - t + t\omega_{\max}|\}$. This expression for the spectral radius has a minimum relative to t when $1 - t + t\omega_{\min}$ and $1 - t + t\omega_{\max}$ are equal in magnitude and opposite in sign. Thus the optimal t is

$$t_{\text{opt}} = \left(1 - \frac{\omega_{\min} + \omega_{\max}}{2}\right)^{-1}.$$

The corresponding spectral radius $(\omega_{\max} - \omega_{\min})/(2 - \omega_{\min} - \omega_{\max})$ is less than 1 even if $t_{\text{opt}} \geq 2$. In practice, the eigenvalues of $dM(\theta^\infty)$ are impossible to predict in advance of knowing θ^∞. For problems with a high proportion of missing data, the value $t = 2$ often works well. For instance, taking $t = 2$ approximately halves the number of iterations until convergence in the Dirichlet and robust regression examples discussed earlier.

The modified algorithm $M_t(\theta)$ also has the desirable property of being locally monotone when $0 < t < 2$.

Proposition 2. Suppose that $\theta^{n+1} = M_t(\theta^n)$ converges to the point θ^∞. Then, for all sufficiently large n, either $\theta^n = \theta^\infty$ or $L(\theta^{n+1}) > L(\theta^n)$.

Proof. Since the increment $\theta^{n+1} - \theta^n$ is given by

$$\theta^{n+1} - \theta^n = -t\,d^{20}Q(\theta^n|\theta^n)^{-1}\,dL(\theta^n),$$

the difference $\Delta(\theta^n) = L(\theta^{n+1}) - L(\theta^n)$ has second-order Taylor expansion

$$\Delta(\theta^n) = dL(\theta^n)^*(\theta^{n+1} - \theta^n) + \frac{1}{2}(\theta^{n+1} - \theta^n)^* d^2L(\phi^n)(\theta^{n+1} - \theta^n)$$

$$= \frac{1}{2}(\theta^{n+1} - \theta^n)^*\left\{d^2L(\phi^n) - \frac{2}{t}d^{20}Q(\theta^n|\theta^n)\right\}(\theta^{n+1} - \theta^n), \tag{8}$$

where ϕ^n is a point on the line segment from θ^n to θ^{n+1}. Now the limit

$$\lim_{n\to\infty}\left\{d^2L(\phi^n) - \frac{2}{t}d^{20}Q(\theta^n|\theta^n)\right\} = d^2L(\theta^\infty) - d^{20}Q(\theta^\infty|\theta^\infty) - \left(\frac{2}{t} - 1\right)d^{20}Q(\theta^\infty|\theta^\infty)$$

is a positive definite matrix because it is expressible as the difference of the positive semidefinite matrix $d^2L(\theta^\infty) - d^{20}Q(\theta^\infty|\theta^\infty)$ and the negative definite matrix $(2/t - 1)\,d^{20}Q(\theta^\infty|\theta^\infty)$. Since the eigenvalues of a matrix depend continuously on its entries (Ortega, 1990), it follows that the quadratic form (8) is positive for n large and $\theta^{n+1} \neq \theta^n$. □

4. GLOBAL CONVERGENCE OF ALGORITHM

We now embark on the more subtle task of investigating global convergence of the EM gradient algorithm. A major impediment to establishing global convergence is the possible failure of the monotonicity property $L(\theta^{n+1}) \geq L(\theta^n)$ far from the maximum point. Although monotonicity appears to be the rule in practice, to establish global convergence in theory we need to enforce monotonicity. From the variety of possible enforcement mechanisms, we elect the natural option of instituting a limited line search at every EM gradient step. This strategy ties in well with certain novel convergence results to be presented here pertaining to any continuous, generalized EM algorithm (Dempster *et al.*, 1977). It is noteworthy that in our two numerical examples monotonicity holds in practice.

Our limited line search involves maximizing $Q(\theta|\theta^n)$ along the EM gradient direction $d(\theta^n) = -d^{20}Q(\theta^n|\theta^n)^{-1}\,dL(\theta^n)$ emanating from the current iterate θ^n. For the line search modified algorithm $A(\theta)$, the next iterate $\theta^{n+1} = A(\theta^n)$ is defined to be the unique point $\theta^n + \alpha^n d(\theta^n)$ maximizing $Q\{\theta^n + \alpha\, d(\theta^n)|\theta^n\}$ for $\alpha \in [0, 1]$. Note that θ^{n+1} exists and is unique because $Q\{\theta^n + \alpha\, d(\theta^n)|\theta^n\}$ is a strictly concave function of α and $d(\theta^n)$ is an ascent direction. When $\theta^n + \alpha\, d(\theta^n)$ is infeasible for some $\alpha \in [0, 1]$, $Q\{\theta^n + \alpha\, d(\theta^n)|\theta^n\}$ decreases as the boundary of the feasible region U is approached owing to the fact that $L(\theta)$ tends to $-\infty$ in this situation. If θ^n is a stationary point of $L(\theta)$, then $d(\theta^n) = 0$ holds, and all α yield $\theta^n + \alpha\, d(\theta^n) = \theta^n$. In this case θ^n is a fixed point of $A(\theta)$. Conversely, any fixed point of $A(\theta)$ is a stationary point of $L(\theta)$. The next proposition establishes the only additional property that is necessary for global convergence.

Proposition 3. The modified EM gradient algorithm $A(\theta)$ is continuous.

Proof. Suppose that some sequence θ^n not necessarily generated by $A(\theta)$ has limit θ^∞. Then $\lim_{n\to\infty}\{d(\theta^n)\} = d(\theta^\infty)$ because $d(\theta)$ is continuous. Now let $\alpha^{n_i} d(\theta^{n_i})$ be a convergent subsequence of the bounded sequence $\alpha^n d(\theta^n)$. Passing to a subsequence if necessary, assume in addition that $\lim_{i\to\infty}(\alpha^{n_i}) = \alpha$. Taking limits on i in the inequality

$$Q\{\theta^{n_i} + \alpha^{n_i} d(\theta^{n_i})|\theta^{n_i}\} \geq Q\{\theta^{n_i} + \beta\, d(\theta^{n_i})|\theta^{n_i}\}$$

produces

$$Q\{\theta^\infty + \alpha\, d(\theta^\infty) \mid \theta^\infty\} \geq Q\{\theta^\infty + \beta\, d(\theta^\infty) \mid \theta^\infty\}$$

for any $\beta \in [0, 1]$. Hence, $\theta^\infty + \alpha\, d(\theta^\infty)$ coincides with the unique optimal point along the direction $d(\theta^\infty)$ emanating from θ^∞. □

From this point onwards, it suffices to assume that $A(\theta)$ is any continuous map of the feasible region U into itself satisfying $L\{A(\theta)\} \geq L(\theta)$, with equality occurring only when θ is a fixed point of $A(\theta)$. The set of fixed points of $A(\theta)$ is assumed to coincide with the set S of stationary points of $L(\theta)$. In the terminology of Dempster *et al.* (1977), $A(\theta)$ is a continuous generalized EM (GEM) algorithm. The original EM algorithm, the ECM algorithm of Meng and Rubin (1993) and our modified EM gradient algorithm all qualify as continuous GEM algorithms. Continuity of the EM and ECM algorithms is almost invariably a consequence of the implicit function theorem. Our first result in this general framework is the basis of all work on discrete Lyapunov functions. Its well-known proof is sufficiently brief to repeat here (Luenberger, 1984).

Proposition 4 (Lyapunov's theorem). Let Γ be the set of limit points generated by the sequence $\theta^{n+1} = A(\theta^n)$ starting from some initial θ^1. Then Γ is contained in the set S of stationary points of $L(\theta)$.

Proof. Consider a typical limit point $\phi = \lim_{k\to\infty}(\theta^{n_k})$. Since $L(\theta^n)$ is monotone and bounded above, $\lim_{n\to\infty}\{L(\theta^n)\}$ exists. Hence, taking limits in the inequality $L\{A(\theta^{n_k})\} \geq L(\theta^{n_k})$ and using the continuity of $A(\theta)$ and $L(\theta)$, we conclude that $L\{A(\phi)\} = L(\phi)$. Thus, ϕ is a fixed point of $A(\theta)$ and consequently also a stationary point of $L(\theta)$. □

The next two propositions are adapted from Meyer (1976).

Proposition 5. The set of limit points Γ of $\theta^{n+1} = A(\theta^n)$ is compact and connected.

Proof. Γ is a closed subset of the compact set $\{\phi \in U\colon L(\phi) \geq L(\theta^1)\}$ and is therefore itself compact. Another theorem of Ostrowski states that Γ is connected provided that $\lim_{n\to\infty}\|\theta^{n+1} - \theta^n\| = 0$ (Ostrowski, 1973). If this sufficient condition fails, then the compactness of $\{\phi \in U\colon L(\phi) \geq L(\theta^1)\}$ makes it possible to extract a subsequence θ^{n_k} such that $\lim_{n\to\infty}(\theta^{n_k}) = \phi$ and $\lim_{n\to\infty}(\theta^{n_k+1}) = \psi$ both exist, but $\psi \neq \phi$. Now the continuity of $A(\theta)$ implies that $\psi = A(\phi)$, and the monotonicity of $A(\theta)$ implies that $L(\psi) = L(\phi) = \lim_{n\to\infty}\{L(\theta^n)\}$. The equality $L(\psi) = L(\phi)$ forces the contradictory conclusion that ϕ is a fixed point of $A(\theta)$. Hence, the sufficient condition $\lim_{n\to\infty}\|\theta^{n+1} - \theta^n\| = 0$ for connectivity holds. □

Proposition 6. Suppose that the feasible region U is open, that the log-likelihood $L(\theta)$ is differentiable and upper compact and that all stationary points are isolated. Then any sequence of iterates $\theta^{n+1} = A(\theta^n)$ generated by a continuous GEM algorithm $A(\theta)$ has a limit, and this limit is a stationary point of $L(\theta)$.

Proof. In the compact set $\{\phi\colon L(\phi) \geq L(\theta^1)\}$ there can only be a finite number of stationary points. Since the set of limit points Γ is a connected subset of this finite set of stationary points, Γ reduces to a single point. □

Some comments about proposition 6 are in order. First, there is no guarantee

that the limit of a GEM sequence furnishes a global maximum; in fact, Wu (1983) gave a counter-example where convergence to a saddlepoint occurs in the EM algorithm. However, any such counter-example is apt to involve unusual symmetries, and convergence to at least a local maximum of the log-likelihood is almost always experienced in numerical practice. Second, Wu (1983) and Boyles (1983) highlighted the importance of the condition $\lim_{n\to\infty}\|\theta^{n+1} - \theta^n\| = 0$ in proving convergence of the EM algorithm. This condition emerges as a consequence of the continuity of $A(\theta)$ in proposition 5. Finally, there is an easily stated sufficient condition that ensures discreteness of the set of stationary points S. Any point θ whose Hessian matrix $d^2L(\theta)$ is non-singular is termed non-degenerate. It is straightforward to verify that every non-degenerate stationary point is isolated (Hestenes, 1981). Thus, S is certainly discrete when all $\theta \in S$ are non-degenerate.

5. DISCUSSION

The essence of the EM algorithm is that it transfers the maximization of $L(\theta)$ to the maximization of the far simpler function $Q(\theta|\phi)$. This maximization transfer perspective suggests how to formulate the EM gradient algorithm in the presence of parameter constraints and bounds. One should attempt to maximize $Q(\theta|\phi)$ subject to the constraints and bounds via the quadratic approximation

$$Q(\theta|\phi) \approx Q(\phi|\phi) + dL(\phi)^*(\theta - \phi) + \tfrac{1}{2}(\theta - \phi)^* d^{20}Q(\phi|\phi)(\theta - \phi).$$

This problem can be solved by standard techniques of quadratic programming.

Parameter constraints and bounds not only complicate implementation, but they also complicate a theoretical development of the algorithm. We have taken the course of defining these complications out of existence. In real examples, parameter estimates do occur on boundaries, and the rate of convergence of the EM algorithm can be sublinear. In spite of these complications, we can often construct a coherent qualitative convergence theory, particularly in examples like medical imaging where the observed log-likelihood is strictly concave (Lange and Carson, 1984).

Even in examples without constraints and boundaries, multiple modes of the likelihood surface are a problem. Caution should be exercised to ensure that the EM gradient algorithm does not converge to an inferior mode. One obvious remedy is to restart the algorithm at a number of different points. It is possible for the EM and EM gradient algorithms started from the same point to converge to different points. In one mixture of normals problem this occurred, with the EM algorithm reaching the better mode. The imposition of a Bayesian prior almost always makes the $Q(\theta|\phi)$ function more concave. We have said little about maximum *a posteriori* analysis for the simple reason that it has almost no effect on how the EM gradient algorithm is formulated or implemented. The log-prior is left untouched by the E-step and is just added to $Q(\theta|\phi)$.

The rate of convergence of the EM gradient algorithm is identical with that of the EM algorithm. In many problems convergence can be frustratingly slow. Doubling the current increment provides some improvement, but more radical measures can be employed. For example, the EM gradient algorithm can be made the basis of a quasi-Newton acceleration (Lange, 1995). On extreme examples, this acceleration technique produces an order of magnitude or more reduction in the number of iterations until convergence.

In conclusion, the ultimate test of any algorithm lies in its performance on a variety of practical problems. The EM gradient algorithm has already proved useful on a few numerical examples. Especially promising are the dramatic improvements seen with the quasi-Newton acceleration. However, there is a clear need for additional testing and rigorous comparison with competing algorithms.

ACKNOWLEDGEMENTS

I thank Tom Belin and Steve Matthysse for suggesting useful improvements to the first draft of my manuscript, Jan de Leeuw for pointing out the reference by Meyer (1976) and the Editor for offering editorial advice. This research was supported in part by the University of California at Los Angeles and US Public Health Service grant CA 16042.

REFERENCES

Boyles, R. A. (1983) On the convergence of the EM algorithm. *J. R. Statist. Soc.* B, **45**, 47–50.

Chiang, C. L. (1980) *An Introduction to Stochastic Processes and Their Applications.* Huntington: Krieger.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

——— (1980) Iteratively reweighted least squares for linear regression when the errors are normal/independent distributed. In *Multivariate Analysis—V* (ed. P. R. Krishnaiah). Amsterdam: North-Holland.

Green, P. J. (1990) On use of the EM algorithm for penalized likelihood estimation. *J. R. Statist. Soc.* B, **52**, 443–452.

Hämmerlin, G. and Hoffmann, K.-H. (1991) *Numerical Mathematics.* New York: Springer.

Hestenes, M. R. (1981) *Optimization Theory: the Finite Dimensional Case.* Huntington: Krieger.

Hille, E. (1959) *Analytic Function Theory*, vol. 1. New York: Blaisdell Ginn.

Lange, K. (1995) A quasi-Newton acceleration of the EM algorithm. *Statist. Sin.*, to be published.

Lange, K. and Carson, R. (1984) EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assist. Tomogr.*, **8**, 306–316.

Lange, K., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modeling using the *t* distribution. *J. Am. Statist. Ass.*, **84**, 881–896.

Lange, K. and Sinsheimer, J. S. (1993) Normal/independent distributions and their applications in robust regression. *J. Comput. Graph. Statist.*, **2**, 175–198.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Luenberger, D. G. (1984) *Linear and Nonlinear Programming*, 2nd edn. Reading: Addison-Wesley.

Meng, X.-L. and Rubin, D. B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Meyer, R. R. (1976) Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *J. Comput. Syst. Sci.*, **12**, 108–121.

Mosimann, J. E. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, **49**, 65–81.

Narayanan, A. (1991) Algorithm AS 266: Maximum likelihood estimation of the parameters of the Dirichlet distribution. *Appl. Statist.*, **40**, 365–374.

Ortega, J. M. (1990) *Numerical Analysis: a Second Course.* Philadelphia: Society for Industrial and Applied Mathematics.

Ostrowski, A. M. (1973) *Solutions of Equations in Euclidean and Banach Spaces.* New York: Academic Press.

Pike, M. C. and Hill, I. D. (1966) Algorithm 291: Logarithm of the gamma function. *Communs Ass. Comput. Mach.*, **9**, 694.

Pratt, J. (1981) Concavity of the log likelihood. *J. Am. Statist. Ass.*, **76**, 103–106.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

Schneider, B. E. (1978) Algorithm AS 121: Trigamma function. *Appl. Statist.*, **27**, 97–99.

Titterington, D. M. (1984) Recursive parameter estimation using incomplete data. *J. R. Statist. Soc.* B, **46**, 257–267.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Wanek, L. A., Goradia, T. M., Elashoff, R. M. and Morton, D. L. (1993) Multi-stage Markov analysis of progressive disease applied to melanoma. *Biometr. J.*, **35**, 967–983.

Wei, G. C. G. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Am. Statist. Ass.*, **85**, 699–704.

Wu, C. F. (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.